DOCUMENT RESUME

ED 302 572 TM 012 584

AUTHOR Ferrara, Steven F.
TITLE Practical Considerations in Equating a Direct Writing Assessment for High School Graduation.
PUB DATE 10 Apr 87
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Essay Tests; *Graduation Requirements; High Schools; High School Students; Interrater Reliability; *Scoring; *Standardized Tests; State Programs; Test Construction; Testing Programs; Test Selection; *Writing Evaluation
IDENTIFIERS Anchor Tests; Direct Assessment; *Maryland Functional Writing Test

ABSTRACT

Statistical equating of scores from direct writing assessments is performed in Maryland to hold the difficulty of passing the Maryland Writing Test (MWT) equivalent from year to year and to allow a demonstration of growth in student writing achievement over the years. Passage of the test is a high school graduation requirement. Prior to statistical analysis, major steps taken in development of equated scores include: (1) prompt development, field testing, and selection; (2) development of anchor, training, qualifying, validity, and check sets; (3) selection, training, and qualifying of raters and application of scoring rubrics by raters; and (4) selection of an equating study sample and selection, administration, scoring, and statistical equating of anchor prompts. Standardization of test administration as well as rater training and scoring, interpretation of adjusted raw scores and scale scores, adaptation of statistical procedures, and sample selection and administration of equating prompts are discussed briefly. Some of the practical problems inherent in equating MWT scores are outlined, and the relationship of these problems to measurement considerations is discussed. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Steven Ferrara
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PRACTICAL CONSIDERATIONS IN EQUATING A DIRECT WRITING ASSESSMENT REQUIRED FOR HIGH SCHOOL GRADUATION

Steven F. Ferrara
Maryland State Department of Education (MSDE)
April 10, 1987

ED302572

In W. Littlefair (Chair), Using a direct writing assessment as a graduation prerequisite: Political, measurement, and instructional issues. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC

Equating scores from direct writing assessments — especially writing assessments like the Maryland Writing Test (MWT) with a passing score and that is required for high school graduation — is necessary to:

1. Provide consistent pass-fail and certification decisions from year to year;

2. Hold the difficulty of passing the test equivalent to that of previous years' tests; and

3. Demonstrate growth in writing achievement over time.

Statistical equating analysis is only the final step in producing equated scores from a direct writing assessment. The major steps that are typically taken to produce equated scores include:

1. Prompt development, field testing, and selection;

2. Development of anchor, training, qualifying, validity, and check sets;

3. Selection, training, and qualifying of raters and application of scoring rubrics by raters; and

4. Selection of an equating study sample and selection, administration, scoring, and statistical equating of anchor prompts.

The goal is to produce equated test scores. The problem is to conduct the equating study while taking into account measurement and practical considerations. These considerations include prompt field testing and selection; standardization of test administration, rater training, and scoring; sample selection and administration of equating prompts; adapting statistical equating procedures; and interpretation of adjusted holistic scores and scale scores. I discuss each of these

TM012584

BEST COPY AVAILABLE

considerations below.

Practical and Other Considerations and MSDE's Current Approaches

Prompt Field Testing and Selection

Field testing and selection of prompts is an essential part of producing a writing assessment that is equivalent in difficulty to previous versions of the test. Preserving the security of prompts is equally important and much more difficult than for selected-response tests in which several field test items can be interspersed with operational items with little risk to item security. Maryland's solution in the past has been to field test new prompts in other states. This strategy assures security but causes concern about the comparability of the out-of-state sample to Maryland students on demographics, writing ability and writing strategies, and motivation of students to perform near their levels of ability.

Our concern for the comparability of data produced from out-of-state field testing has led us to conduct in-state field tests this year. In-state field testing involves some risks to prompt security: as much as one-third of the field test sample students is likely to have failed the MWT and some small portion of these students will encounter the field tested prompts again next year as operational prompts. In addition, and despite our efforts, some students may have managed to copy down or memorize field test prompts and to share them with other students. However, we believe that the improvement in the comparability of this year's field test data to operational test data outweighs these small risks.

MSDE is currently making plans to calibrate and scale writing prompts using the polychotmous form of the Rasch model, referred to as the Partial Credit model (Wright & Masters, 1982). Our hope is that model-data fit will be adequate so that we can bank prompts and build future editions of the MWT from biannual field tests.

Standardization of Test Administration, Rater Training, and Scoring

Activities like these usually are conducted by writing specialists rather than measurement specialists. Although these activities are not typically discussed in the context of test score equating they are an essential part of the process of equating direct writing assessments. The degree to which these activities are standardized and completed with the same levels of rigor as in previous years determines in part the equivalence of scores in a given year to previous years. Any ways in which the standardization of these activities differs from previous years can contribute to non-equivalence and necessitate score adjustments. Practical measures for assuring the quality of rater training and scoring are widely known and applied in the MWT program. Some of these measures are discussed by other presenters in the symposium.

Sample Selection and Administration of Equating Prompts

Conducting an annual equating study has always been a source of difficulty for both MSDE and local school systems. Obtaining data to conduct an equating analysis requires that students in several schools around the state participate in the regular administration of the MWT and in a special administration of "equating" prompts (from a previous year's test). This additional test administration disrupts schools for two half-days and is seen as a somewhat onerous task by students and school staff. To minimize disruptions to schools we ask 10-12 schools in 5-6 districts to participate. Districts and schools are selected to produce a sample representative of the state on key variables and with consideration for regular school activities that cannot be disrupted (e.g., semester exams) and participation in other MSDE assessment activities like the in-state field test. MSDE considers participation in the annual equating study the duty of local schools and districts but also an important service in support of MSDE efforts.

Adapting Statistical Procedures

Tried-and-true statistical procedures for equating test scores (e.g., linear and equipercentile equating) are designed for tests with large numbers of selected-response items. Such tests allow adjustments in scores as small as one item. However, direct writing assessments are typically configured as one or a pair of essays or one essay with a number of selected-response items. Such configurations usually produce a total test score scale with a limited number of score points on which to conduct an equating analysis and make necessary score adjustments.

The MWT, which allows adjacent agreements between two raters and consists of two essays, produces a total test scale with 17 score points (0.0, 1.0, 1.5,...7.5, 8.0). This scale allows us to adjust raw (i.e., holistic) scores — and specifically the passing raw score — only if the difference in scores for the two equated tests is at least 0.5. Table 1 below contains equated raw and scale scores from the 1986 MWT.

Table 1

Equated Raw and Scale Scores From the 1986 MWT (Portion of Scale Around the Passing Score)

Observed 1984 Raw Score	Equated 1986 Raw Score	Equated 1986 Scale Score
6.0	5.51	345
5.5	5.13	340
5.0	4.82	337

Note . Passing score is 5.5

The data in Table 1 above demonstrate that there was some difference between the 1986 and anchor year passing scores. However, this difference was not great enough (i.e., $\geq$ 0.5) to warrant any adjustment to scores.

Interpretation of Adjusted Raw Scores and Scale Scores

Typically, if an equating analysis for a selected-response test indicates that a current year's test is more difficult or easier than previous years' tests, raw score to scale score equivalences are adjusted. For selected-response tests such a difference may be reflected as a one or two raw score point difference in the test's passing score. Such a result from an equating analysis for the MWT may seem to suggest that the total test passing score should be adjusted up to 6.0 or down to 5.0 (as in Table 1 above) from the current passing score of 5.5. However, this adjustment may communicate to the public that the passing standard is changing from year to year, rather than being adjusted to account for yearly differences in test difficulty. Another solution, with its own drawbacks, is to introduce total test scale scores that mask the adjustments made to raw scores.

One drawback to reporting total test scale scores occurs the first time they are reported: students, parents, and teachers are unlikely to know how to interpret them, especially if they are accustomed to interpreting total test holistic scores. Scale score scales do take on meanings of their own over time (take, for example, the SAT scale), but only after several years of practice at associating score levels with student performance levels on other measures and even with specific students. A simple solution is to report scale scores along with total test holistic scores and train consumers in their interpretation prior to the year when passing score adjustment becomes necessary.

A second drawback related to using scale scores is a problem not with interpreting scale scores themselves but with interpreting individual essay scores after an adjustment has been made to total test scores. Since the score equating is necessarily accomplished on the total test score scale interpretation of individual essay scores is no longer clear. Using the MWT as an example, if an equating analysis indicates that a current year's test in more difficult in the past then an adjustment on the total test scale is made is to reflect this difference. However, the meaning of the holistic scores on each individual essay is unclear: Does a score of 3.0 on the narrative essay mean the same as in previous years? What about a 3.0 on the explanatory essay? Should these two scores be interpreted more like a 3.5 in previous years? Or should such an interpretation be made for only one of the two essays? Which one? Other information can be brought to bear (e.g., average scores on the two essays on the current

and anchor form of the test), but this information allows only educated guess work. Further, such guess work is likely to be unpalatable to students, parents, and teachers.

Other solutions are available but, again, have their drawbacks. One such solution would be to equate individual essays. However, such an equating is inadvisable because even fewer score points are available for equating and adjusting scores. Another, more complex and sophisticated approach is exemplified by the Levels of Proficiency scale used to report performance on the National Assessment of Educational Progress 1985 reading assessment (Educational Testing Service, 1986). To create the scale, national reading experts associated reading behaviors with clusters of similarly difficult and discriminating items within ranges of scale scores. Descriptive labels for each scale score range were then derived from the reading behaviors associated with the cluster of items and score range. As a result, NAEP reading scores of around 300 indicate "adept" reading proficiency. This solution to aiding interpretation of equated scale scores for a writing assessment seems viable but requires a considerable amount of time and development effort.

MSDE currently reports only holistic scores for each essay and total test holistic scores. We occasionally remind school district staff that if an equating analysis indicates the need for (passing) score adjustment, this adjustment would be made via scale scores. Beginning in 1989 when we administer multiple-choice writing conventions items, the only total test score we will report will be a scale score.

Conclusion

We believe in Maryland that statistical equating of writing assessments is necessary both to hold the difficulty of passing the MWT equivalent from year to year and to allow us to demonstrate growth in student writing achievement over the years. I have outlined some of the practical problems inherent in equating MWT scores and the relationship of these problems to measurement considerations. Although current solutions to these problems may not seem ideal, they are adequate and seem to be working for the MWT.

References

Educational Testing Service. (1986). The reading report card. (Report No. 15-R-Ø1). Princeton, NJ: Author.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: Mesa Press.